Reconstruction of Interaction Networks (With Applications to Transcriptional Regulation)

Ilya Nemenman
JCSB, Columbia

Andrea Califano, Adam Margolin, Chris Wiggins, Gustavo Stolovitzky, Kai Wang, Riccardo Dalla Favera, Katia Basso, Ulf Klein, Nila Banerjee

q-bio.MN/0411003, q-bio.MN/0410037
q-bio.MN/0410036, q-bio.QM/0406015
Gene expression analysis
Gene expression analysis

- clustering – too coarse
- reconstructing networks – Holy Grail!
Gene expression analysis

- clustering – too coarse
- reconstructing networks – Holy Grail!

Many methods exist, but:
- loops?
- what dependence (arrows) means?
- what approximations being made? controlling them?
- are approximations biologically sound?
- guarantees?
Gene expression analysis

- clustering – too coarse
- reconstructing networks – Holy Grail!

Many methods exist, but:
- loops?
- what dependence (arrows) means?
- what approximations being made? controlling them?
- are approximations biologically sound?
- guarantees?

Different conditions – different steady states.
Gene expression analysis

- clustering – too coarse
- reconstructing networks – Holy Grail!

Many methods exist, but:
- loops?
- what dependence (arrows) means?
- what approximations being made? controlling them?
- are approximations biologically sound?
- guarantees?

Different conditions – different steady states.
Extremely many false positives (e.g. joint co-regulation).
Model of dependence

No time series → steady state statistical dependencies only.
Model of dependence

No time series $\rightarrow$ steady state statistical dependencies only.

$$-\log P(g_i) = \sum_i \phi_i(g_i) + \sum_{ij} \phi_{ij}(g_i, g_j) + \sum_{ijk} \phi_{ijk}(g_i, g_j, g_k) + \ldots$$
Model of dependence

No time series $\rightarrow$ steady state statistical dependencies only.

$$-\log P(g_i) = \sum_i \phi_i(g_i) + \sum_{ij} \phi_{ij}(g_i, g_j) + \sum_{ijk} \phi_{ijk}(g_i, g_j, g_k) + \ldots$$

- use MaxEnt to define $\phi$
- connections with spin glasses, MNs, belief propagation
Model of dependence

No time series $\rightarrow$ steady state statistical dependencies only.

$$-\log P(g_i) = \sum_i \phi_i(g_i) + \sum_{ij} \phi_{ij}(g_i, g_j) + \sum_{ijk} \phi_{ijk}(g_i, g_j, g_k) + \ldots$$

- use MaxEnt to define $\phi$
- connections with spin glasses, MNs, belief propagation
- enough data to evaluate 2-way marginals only;
Model of dependence

No time series $\rightarrow$ steady state statistical dependencies only.

$$-\log P(g_i) = \sum_i \phi_i(g_i) + \sum_{ij} \phi_{ij}(g_i, g_j) + \sum_{ijk} \phi_{ijk}(g_i, g_j, g_k) + \ldots$$

- use MaxEnt to define $\phi$
- connections with spin glasses, MNs, belief propagation
- enough data to evaluate 2-way marginals only;
- truncate at 2nd order potential (cannot reconstruct XOR), Bethe approximation (but inverse problem)
- Mutual information $I(g_i, g_j) = I_{ij}$ is enough to establish dependencies.
Notes

• changing \( \phi(g_i) \) describes response to perturbations (but: directionality)
Notes

• changing $\phi(g_i)$ describes response to perturbations (but: \textit{directionality})
• biochemical dependencies persist as steady state statistical dependencies, but orders of interactions may change
Removing false positives – Data Processing inequality

\[ I(A, C) \leq \min[I(A, B), I(B, C)] \]
Removing false positives – Data Processing inequality

\[ I(A, C) \leq \min[I(A, B), I(B, C)] \]

**ARACNE:** Look at every triplet and remove the weakest link.
Removing false positives – Data Processing inequality

\[ I(A, C) \leq \min[I(A, B), I(B, C)] \]

**ARACNE:** Look at every triplet and remove the weakest link. Every 3-gene loop is opened!
Guarantees

**Theorem.** If MIs can be estimated with no errors, then ARACNE reconstructs the underlying interaction network exactly, provided this network is a tree and has only pairwise interactions.
Theorem. The maximum Mutual Information spanning tree (Chow-Liu) is a subnetwork of the network reconstructed by ARACNE.
Theorem. Let $\pi_{ik}$ be the shortest path between $i$ and $k$. Then, if MIs are known, ARACNE reconstructs an interaction network without false positives edges, provided: (a) the network consists only of pairwise interactions, (b) for each $j \in \pi_{ik}$, $I_{ij} \geq I_{ik}$. 
Theorem. Let $\pi_{ik}$ be the shortest path between $i$ and $k$. Then, if MIs are known, ARACNE reconstructs an interaction network without false positives edges, provided: (a) the network consists only of pairwise interactions, (b) for each $j \in \pi_{ik}$, $I_{ij} \geq I_{ik}$. 

$$I_{AC} \geq I_{AD}$$

$$I_{AB} \geq I_{AD}$$
Why should it work?

- higher order interactions project into lower order ones
- large loops are locally trees (biological signals decorrelate very fast: $I(c\text{MYK}, c\text{MYK}) \approx 8$ bits, $I(c\text{MYK}, \text{second best}) \approx 1$ bit.
- small loops (e.g., feed forward) are often transient
Kernel MI estimation: Copula transform
Kernel MI estimation: Copula transform
Kernel MI estimation: Copula transform

Do before estimating MI. No need for spatial inhomogeneity.
Mutual information error vs. ranking error
Mutual information error vs. ranking error

Can use universal best $h$. 
Synthetic networks

a

b
Synthetic networks

\[ \frac{dx_i}{dt} = a_i \prod_j \left( \prod_j I_{I_j, j}^{\nu_j} I_{I_j, 0}^{\nu_j} \prod_j \left( 1 + \frac{A_{I_j}^{\nu_j}}{A_{I_j}^{\nu_j} + A_{I_j, 0}^{\nu_j}} \right) \right) - b_i x_i \]
Benchmarks

\[ N_{TP} - N_{FP} = \max \text{ at } p = 10^{-4}. \]
No sampling catastrophe!
c-MYC TF centered network

Protooncogene, involved in many cellular processes, 12% background interactions, top 5% genetic hub, significant MI with $\sim 2000$ genes.
Protooncogene, involved in many cellular processes, 12% background interactions, top 5% genetic hub, significant MI with $\sim 2000$ genes.

- 56 1st neighbors
- pre-known targets – 22
- ChIp-proven targets – 11/12
- 2nd neighbors – weaker enrichments
- Most 1st – major hubs
Hub 3-way interactions (conditional analysis)

\[ I(g_i, g_j | G_{\mu}^*) \]

- \( G_{\mu}^* \) – coarse conditions (+/-) of correlated gene clusters
- Independent of the hub (true 3-way interactions)
- Large dynamic range

<table>
<thead>
<tr>
<th></th>
<th>( G_1^+ )</th>
<th>( G_1^- )</th>
<th>( G_2^+ )</th>
<th>( G_2^- )</th>
<th>( \ldots )</th>
<th>( \ldots )</th>
<th>( G_M^+ )</th>
<th>( G_M^- )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Edge 1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Edge 2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Edge N</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Edge support conditions set size

![Graph showing edge support conditions set size]

<table>
<thead>
<tr>
<th>#</th>
<th>$N_p$</th>
<th>$N_v$</th>
<th>$E$</th>
<th>$N_{FP}$</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2422</td>
<td>437</td>
<td>0.18</td>
<td>6520.1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1458</td>
<td>278</td>
<td>0.19</td>
<td>4541.5</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1066</td>
<td>224</td>
<td>0.21</td>
<td>2514.1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>847</td>
<td>182</td>
<td>0.21</td>
<td>1131.6</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>710</td>
<td>157</td>
<td>0.22</td>
<td>423.13</td>
<td>0.60</td>
</tr>
<tr>
<td>6</td>
<td>591</td>
<td>132</td>
<td>0.22</td>
<td>136.04</td>
<td>0.23</td>
</tr>
<tr>
<td>7</td>
<td>511</td>
<td>119</td>
<td>0.23</td>
<td>37.03</td>
<td>0.072</td>
</tr>
<tr>
<td>8</td>
<td>459</td>
<td>110</td>
<td>0.24</td>
<td>9.18</td>
<td>0.02</td>
</tr>
<tr>
<td>9</td>
<td>406</td>
<td>104</td>
<td>0.26</td>
<td>1.9</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>10</td>
<td>367</td>
<td>96</td>
<td>0.26</td>
<td>0.37</td>
<td>&lt; 0.01</td>
</tr>
</tbody>
</table>
Edge support conditions set size

Probably better than original algorithm.
Conditional network sizes
Regulators, indeed

Of 168 c-MYC regulators:

<table>
<thead>
<tr>
<th>GO Category</th>
<th>$N_C$</th>
<th>$N_{Tot}$</th>
<th>GO $N_C$</th>
<th>$\frac{GO}{N_{Tot}}$</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transcription Regulator Activity (MF)</td>
<td>21</td>
<td>116</td>
<td>2089</td>
<td>21014</td>
<td>0.0049</td>
</tr>
<tr>
<td>Protein Kinase (MF)</td>
<td>13</td>
<td>116</td>
<td>1004</td>
<td>21014</td>
<td>0.003</td>
</tr>
<tr>
<td>IKBK/NFKB cascade (BP)</td>
<td>5</td>
<td>117</td>
<td>222</td>
<td>24373</td>
<td>0.004</td>
</tr>
<tr>
<td>Immune Response (BP)</td>
<td>19</td>
<td>117</td>
<td>1664</td>
<td>24373</td>
<td>0.0003</td>
</tr>
<tr>
<td>Humoral IR (BP)</td>
<td>9</td>
<td>117</td>
<td>378</td>
<td>24373</td>
<td>0.0001</td>
</tr>
<tr>
<td>Reg. of Transcription, DNA-dep. (BP)</td>
<td>26</td>
<td>117</td>
<td>1697</td>
<td>24373</td>
<td>1.2x10^{-7}</td>
</tr>
</tbody>
</table>